

Deep Learning für Automatische Dokumentanalyse

apl. Prof. Marcus Liwicki

DIVA Group, University of Fribourg

MindGarage, University of Kaiserslautern

marcus.liwicki@unifr.ch

www.mindgarage.de

Können Sie diese Symbole erkennen?

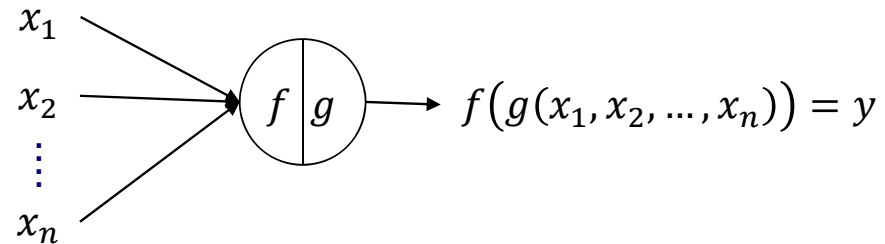
B

Bebop

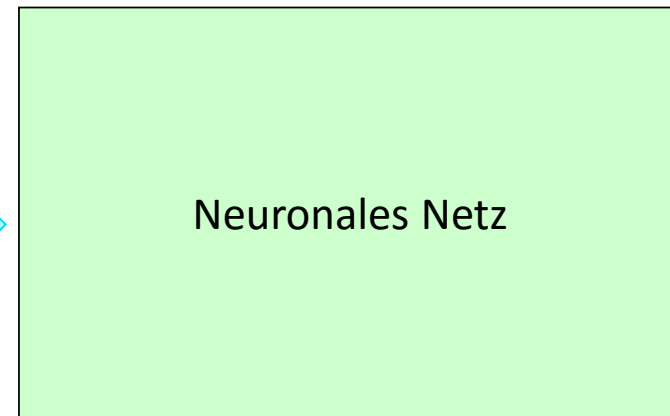
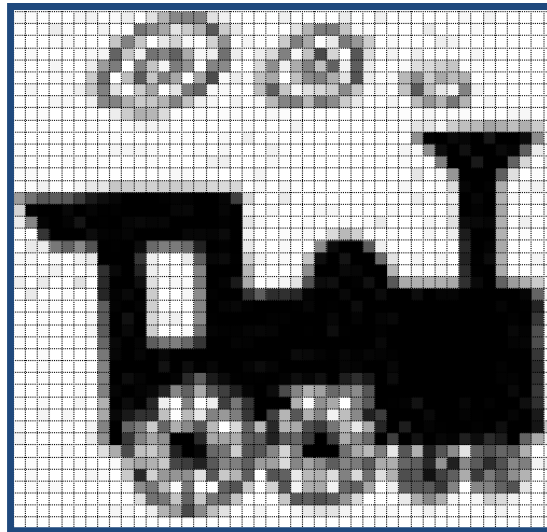
Der Bebop ist eine Musikrichtung, die Anfang der 1940er Jahre den Ursprung des Modern Jazz bildete.



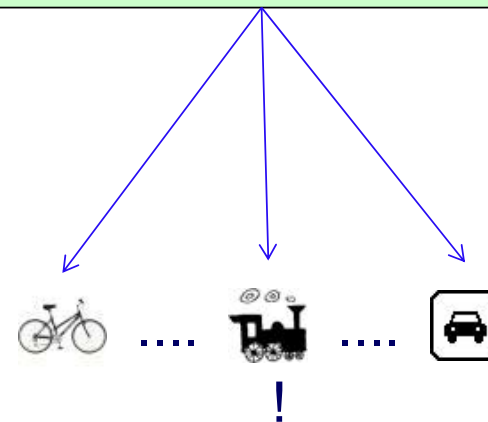
>10¹¹ Neuronen
Idee: Mit Computer simulieren



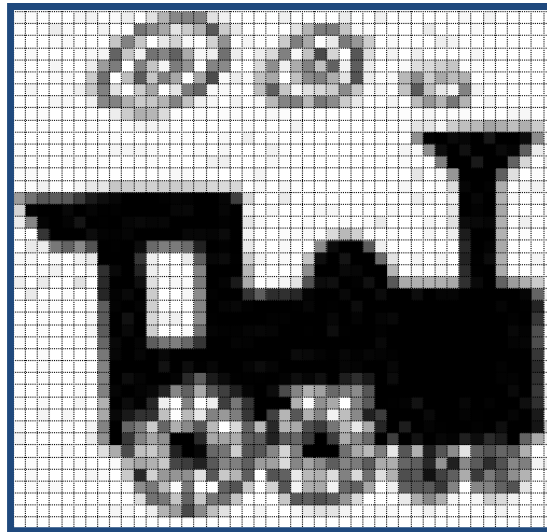
Konzept der Objekterkennung



■ Bild eines Objektes



Konzept des Multi-Layer Perceptrons (MLP)

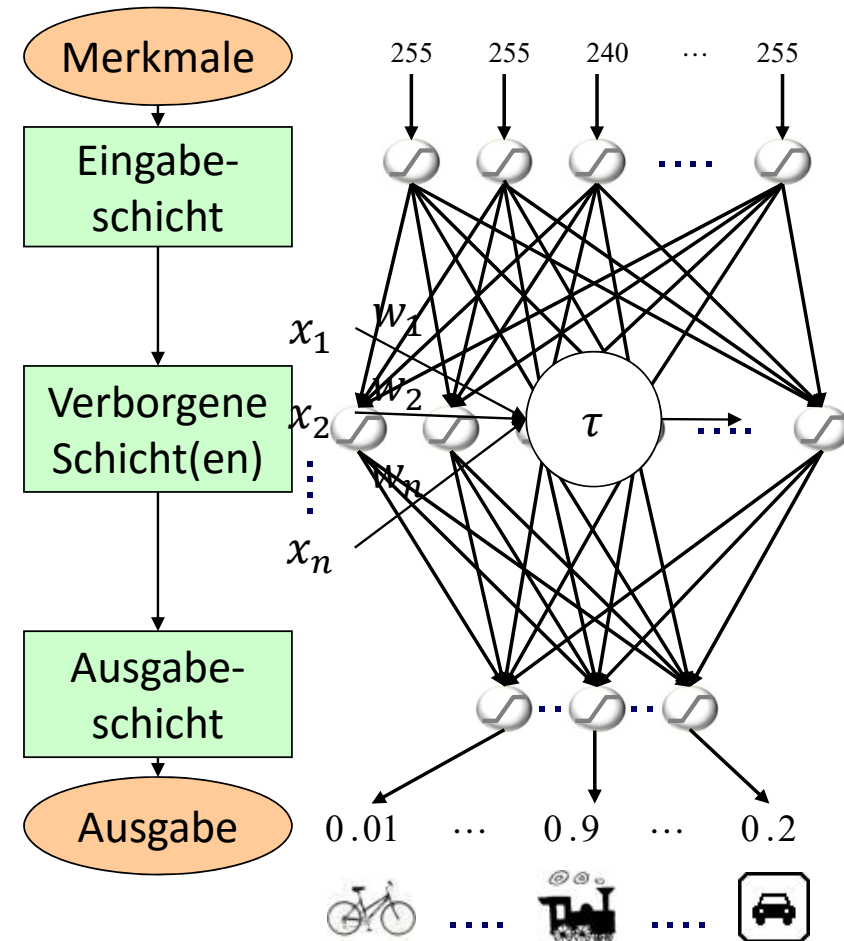


- Bild eines Objektes

255	255	240	...	255
255	248	232	...	255

- Merkmale zur Repräsentation

252	247	238	...	239
⋮	⋮	⋮	⋮	⋮
255	255	255	...	255



Können Sie diese Symbole erkennen?

einfach

B

Bebop

Der Bebop ist eine Musikrichtung, die Anfang der 1940er Jahre den Ursprung des Modern Jazz bildete.

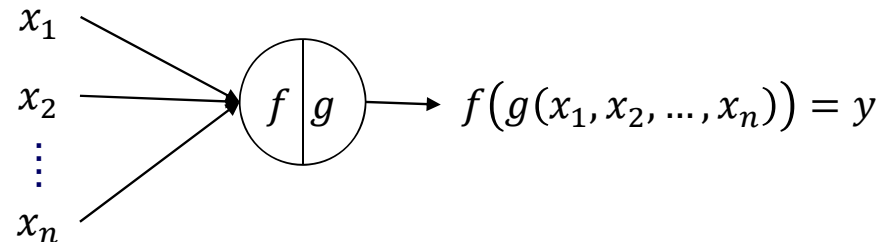
komplex

→ tiefere Strukturen (Deep Learning)



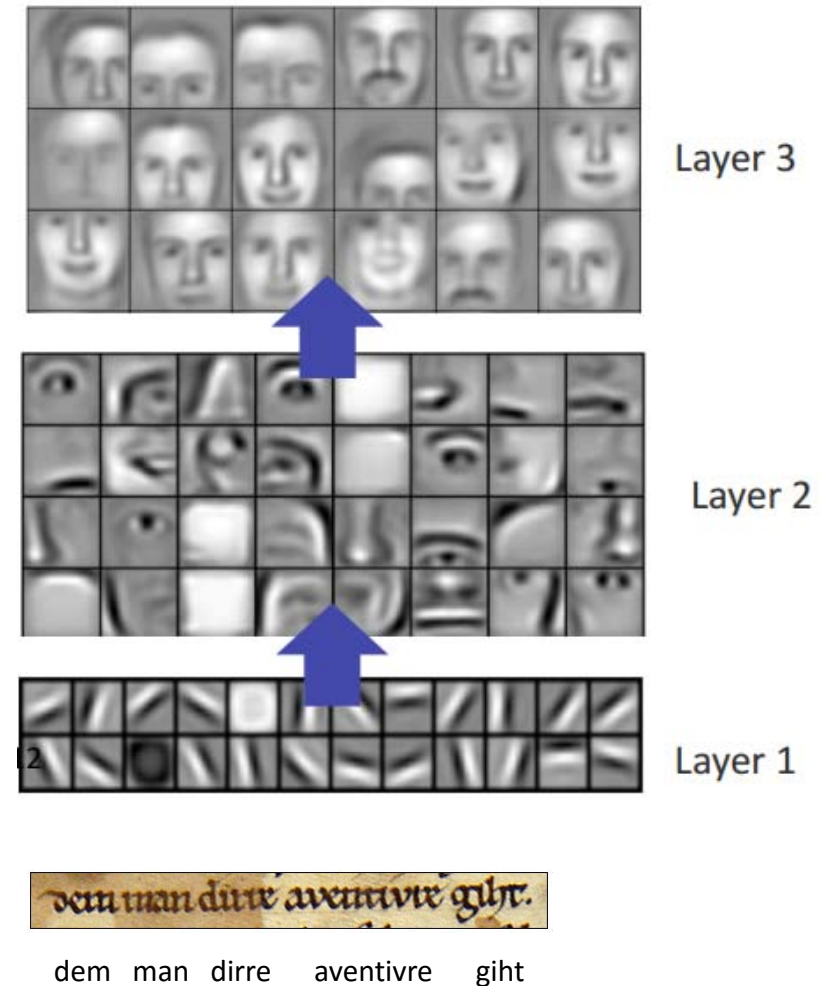
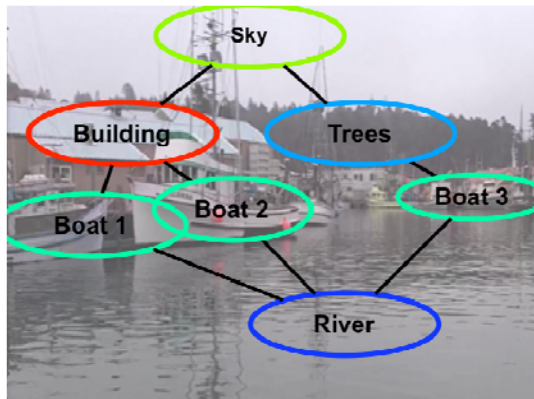
>10¹¹ Neuronen

Idee: Mit Computer simulieren



Mind Garage Was ist Deep Learning

- Sehr tiefe Neuronale Netze
 - ^ Viele Schichten (CNN)
 - ^ Lange Sequenzen (LSTM)
- Lernen von Repräsentationen
- Lernen von Konzepten und Zusammenhängen



Deep Learning für Pixel-Labeling



Parzival (Cod. 857, page 144,
Abbey Library of St. Gall, (PAR23))

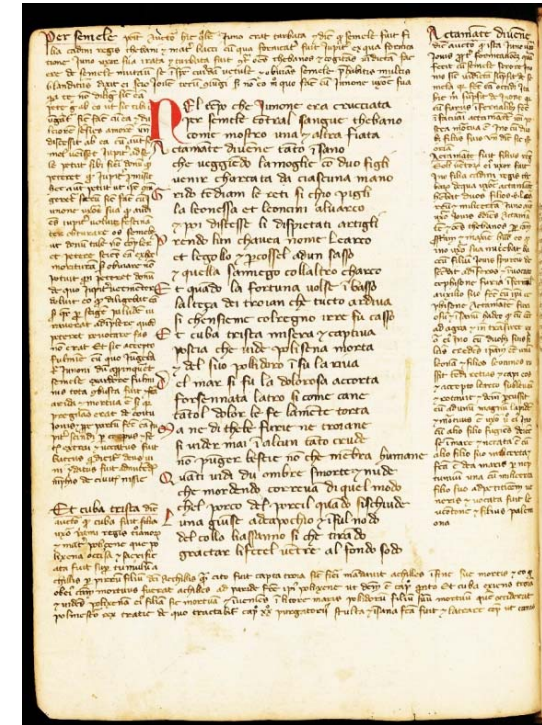
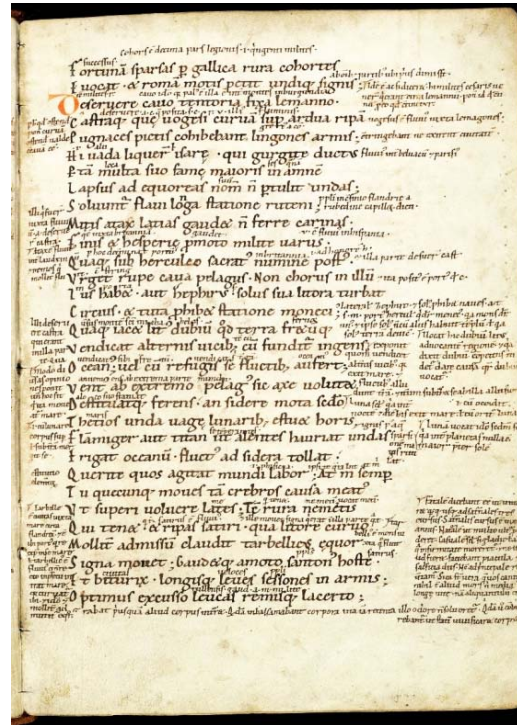
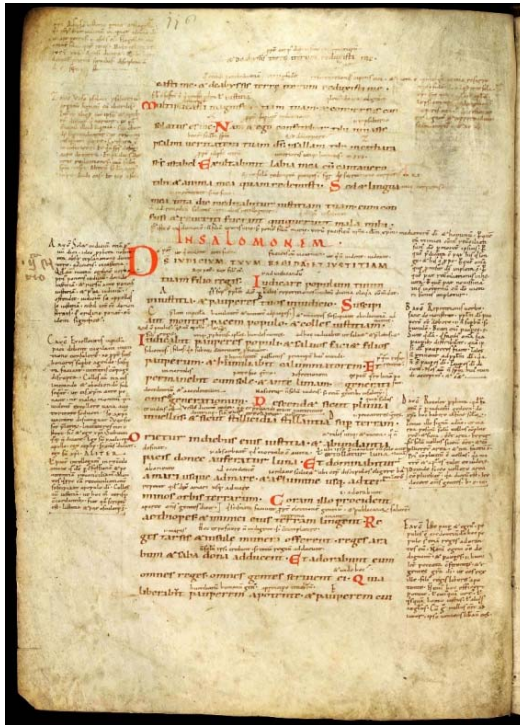


Ground Truth

z. B. Layout-Kategorien:

- Text
- Decoration
- Background

Fischer, Andreas, et al. "Ground truth creation for handwriting recognition in historical documents." Proceedings of the 9th IAPR International Workshop on Document Analysis Systems. ACM, 2010.



(a)

(a) St. Gallen, Stiftsbibliothek, Cod. Sang. 18, (CSG18), Latin, 10th cent.

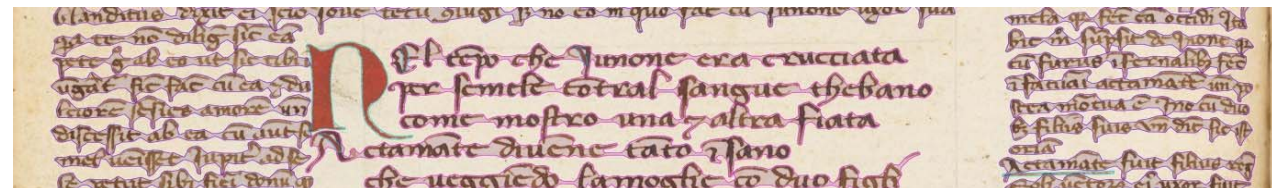
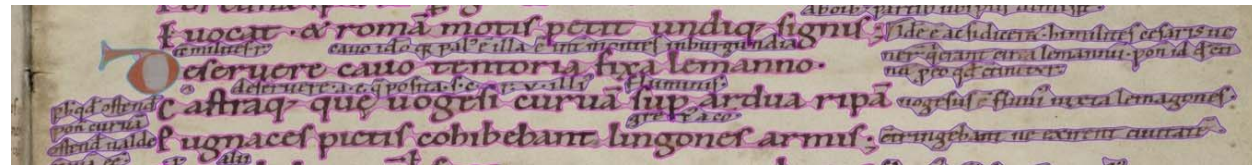
(b)

(b) St. Gallen, Stiftsbibliothek, Cod. Sang. 863 (CSG863), Latin, 11th cent.

(c)

(c) Cologny-Geneve, Fondation Martin Bodmer, Cod. Bodmer 55 (CB55), Italian/Latin glosses, 14th cent.

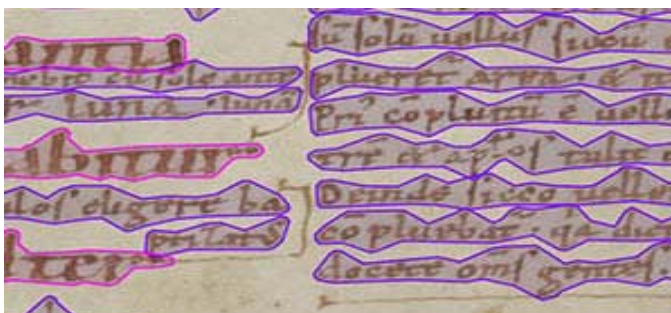
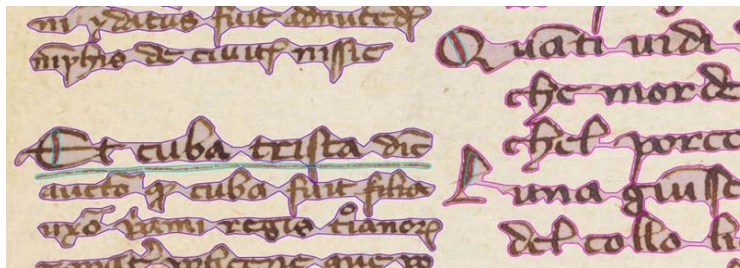
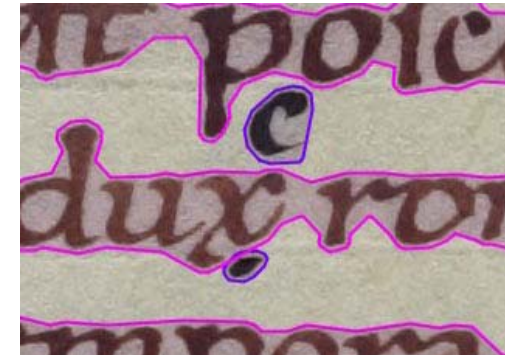
Genauer GroundTruth & Multi-Labeling



decorations

comments

main text body

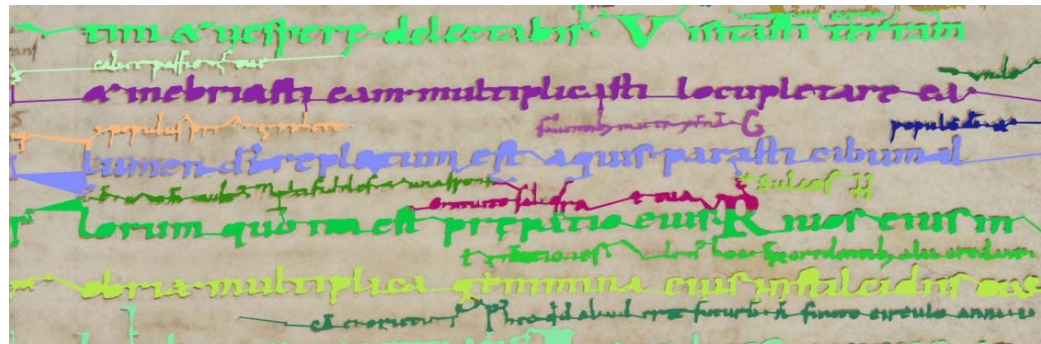


Number of Items per Annotated Category

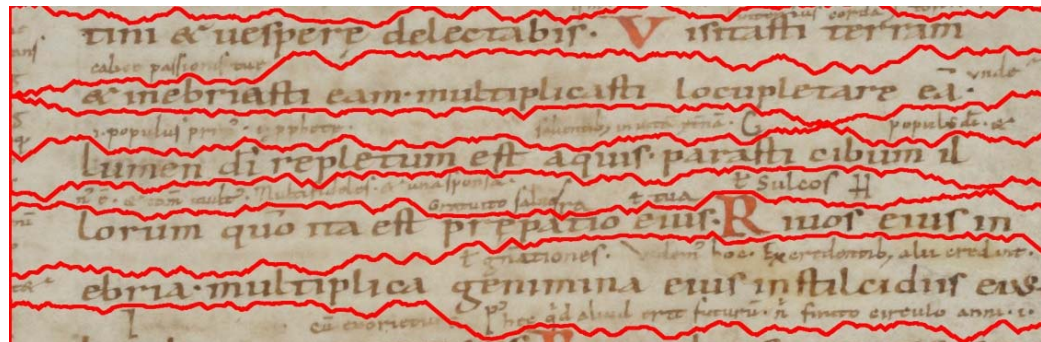
	CSG18	CSG863	CB55	total/MS	(%)	
main text body	1,353	1,538	1,486	4,377	26.67	29.41
decorations	672	30	835	1,537	9.36	0.49
comments	6,260	1,656	2,584	10,500	63.97	70.10
total	8,258	3,224	4,905	16,414		

Textzeilenerkennung ist schwierig

OCRopus



Seam carving



DIVAServices
were used

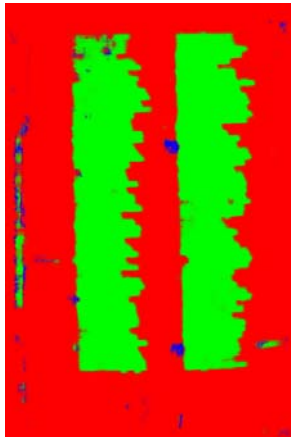
Pixel-Level (in %)	CSG18	CSG863	CB55
background	91.56	93.40	96.05
main text body	94.22	92.08	93.50
decorations	81.58	81.66	93.68
comments	91.35	99.91	90.02
average	89.93	91.76	93.31



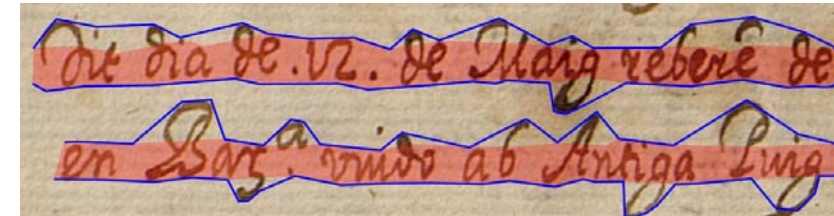
<https://diuf.unifr.ch/main/hisdoc/diva-hisdb>

Deep Learning zur Textzeilenerkennung

Input Image

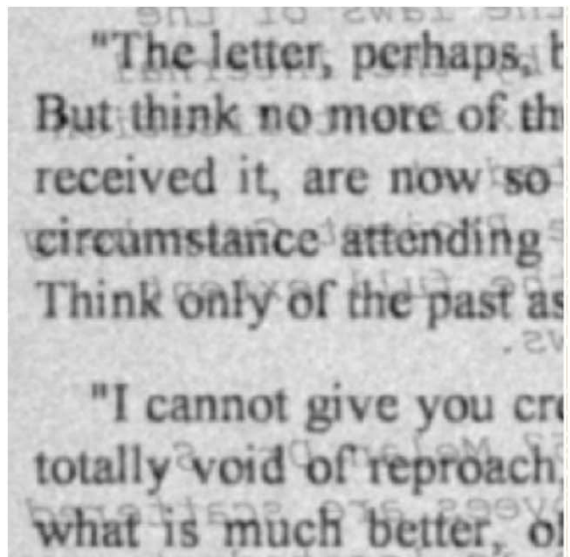
Layout Analyse
(CNN)Main Body Area
(CNN)Segmentation
(watershed
algorithm)

Finally: post-processing (for ascenders/descenders)

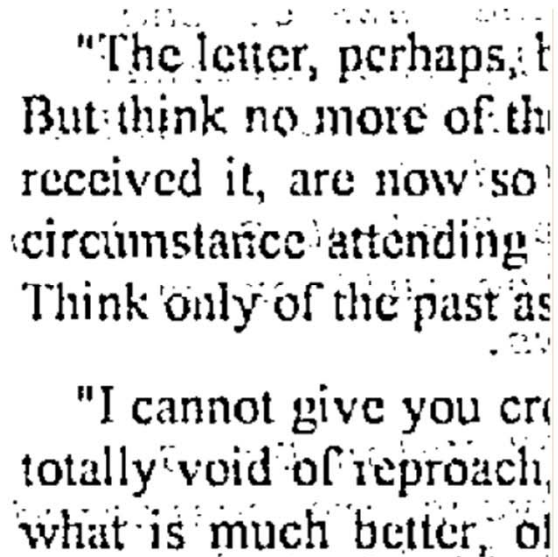


Pastor-pellicer, J., Afzal, M. Z., & Liwicki, M. (2016). Complete Text Line Extraction with Convolutional Neural Networks and Watershed Transform. In 12th IAPR Workshop on Document Analysis Systems (pp. 30–35).

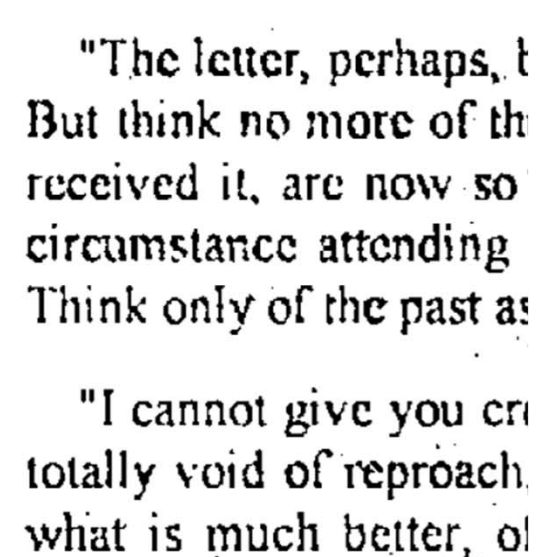
Original



Sauvola



LSTM



OCR from 43% to 73%

Afzal, M. Z., et. al (2015). Document Image Binarization using LSTM : A Sequence Learning Approach. 3rd Int. Workshop on Historical Document Imaging and Processing (pp. 79–84).

Layout Analysis on Modern Documents

■ Very good results using LSTM



Fig. 1. Original LHD image with the cutlines on the task table

In figure of the table when their distance from the first horizontal cutline is less than the distance from the second horizontal cutline, the first cutline is considered as the ground truth. For a particular time in future, only one of these conditions will have a value of one to indicate whether the time to failure is fact, second, or not.

When the time to failure is more than the second type of input time, the value of the status of the input time is considered as one. In the task table, the value of the input time is considered as one. In the task table, the value of the input time is considered as one. In the task table, the value of the input time is considered as one.

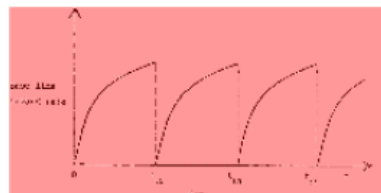


Fig. 2. Graph of results based on open binary output



Fig. 1. Original LHD image with the cutlines on the task table

In figure of the table when their distance from the first horizontal cutline is less than the distance from the second horizontal cutline, the first cutline is considered as the ground truth. For a particular time in future, only one of these conditions will have a value of one to indicate whether the time to failure is fact, second, or not.

When the time to failure is more than the second type of input time, the value of the status of the input time is considered as one. In the task table, the value of the input time is considered as one. In the task table, the value of the input time is considered as one.

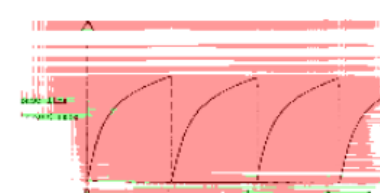


Fig. 2. Graph of results based on open binary output

➔ What is Ground Truth???

OCR Error Correction using LSTM

- Given several OCR (Tesseract and ocropy), post process the errors as much as possible

Dataset	OCR1	OCR2	ISRI[9]	Pairwise[5]	Line-Page	LSTM
English	2.0%	1.56%	1.45%	1.32%	1.26%	0.40%
Fraktur	2.7%	2.5%	2.41%	2.36%	2.31%	0.39%

- Results on public English Dataset and Fraktur: Fontane “Wanderung durch die Mark Brandenburg” (1862-1889)

Azawi, M. Al, Liwicki, M., & Breuel, T. M. (2015). Combination of Multiple Aligned Recognition Outputs using WFST and LSTM. In *13th International Conference on Document Analysis and Recognition* (pp. 31–35). IEEE.

- Deep Learning wird bereits verwendet (SmartOCR)
- Beispiel von jüngeren Drucken
- Fast fehlerfreie OCR
- Mögliche Direkteinbindung in Datenbanken durch Skripte

Globale Extraktionsparameter: Filetemp\HistDocImages\230-6_B_068_0037.jpg

Dividenden auf Stammaktien:
 II/1948/49-1968: insgesamt 237 %
 1969-1973: je 15 % (Div. Sch. Nr. 23-27)

Aus den Bilanzen (in 1 000 DM)

	1972 ¹⁾	1973
Anlagevermögen	85 930	86 678
(Beteiligungen)	4 250	4 337
Umlaufvermögen	163 510	231 222
(flüssige Mittel)	8 260	8 690
Eigenkapital	88 550	115 145
(Grundkapital)	43 200	43 200
Fremdkapital	154 690	201 138
Bilanzgewinn	5 220	6 480
Bilanzsumme	249 450	317 901

Aus den Gewinn- und Verlustrechnungen

	1972 ¹⁾	1973
Umsatzerlöse	441 750	638 702
Beteiligungserträge	2 630	870
Materialaufwand	233 490	337 590
Personalaufwand	91 870	142 026
Abschreibungen	12 080	16 263
EEV-Steuer	21 230	17 280
Jahresüberschuß	10 420	12 960

1) nach Übernahme der Vereinigte Draht- und Kabelwerke AG und der Deutsche Kabelwerke AG

AESCULAP-WERKE AKTIENGESELLSCHAFT
 vormals Jetter & Scheerer

Sitz: 7200 Tuttlingen, Möhringer Straße 125, Postfach 40
Fernruf: (0 74 61) 60 21
Fernschreiber: 7 62 608

Werke: 10 Fabrikgebäude mit Nebengebäuden
Beteiligungen:
 Industrias Quirurgicas Reunidas S.A., Motrico (Spanien)
Kapital: Ptas 3 500 000,- (100 %)

Verifizieren: Bildstapel Dokument Feld Tabelle Ansicht Extras Hilfe

Bildstapel Dokument Feld Tabelle Ansicht Extras Hilfe

Bildbereich

Klassifikation

Klasse (Alle): Daheim/Default

Detailbereich

3. Preuß. Postillon zur Zeit Friedrich II.

Vorname
Friedrich

Arbeitsbereich

Orte	Ort
1	Berlin
2	Düsseldorf
3	Merseburg
4	Hagen
5	Halle
6	

Vornamen	Vorname
1	Elisabeth
2	Friedrich
3	

Verifizieren: Bildstapel verifizieren [DM_Uni Würzburg]

Bildstapel Dokument Feld Tabelle Ansicht Extras Hilfe

Bildleiste Bildbereich

Klassifikation

Klasse (Alle): Daheim/Default

Detailbereich

Vorname

Arbeitsbereich

Orte

	Ort
1	Berlin
2	Düsseldorf
3	Merseburg
4	Hagen
5	Halle
6	

Vornamen

	Vorname
1	Elisabeth
2	Friedrich
3	

Vorgang NONE Dokument 0000 Seite 1/1 Bild 1/1 DE-20_54_A_101_2077-10_27-52_0314_0826_J8H_1874_10_52.pdf:0 VER

826

haupt in jener Zeit das Reisen vor sich ging, selbst wenn es sich um hohe Personen handelte, mag die Tatsache bekunden, daß z. B. die Königin von Polen 1733 zur Zurücklegung einer Strecke von 60 Meilen 11 Tage benötigte, wobei besondere Postkutsche „zum Beleuchten“ des in trostlosestem Naturzustande befindlichen Weges verwendet werden mußten. Und selbst noch kurz vor Erfindung der Schnellposten war ja in des heiligen römischen Reichs Streusandbüchse, der Mark Brandenburg, das Reisen auf Sandwegen von so urwüchsiger Romantik, daß eine Postkutschawane jener Zeit, ein Monstrum von „Landkutsche“ mit abgetriebenen Rössen, daneben 4—6 Passagiere in tiefem Wüstenlande watend, natürlich bei 25° Reaumur, ohne weiteres einem Hogarth als Vorwurf zu einem Genrebilde dienen konnte.

Die englische Mail-Coach, von Mr. Palmer, dem Bristol Theaterdirektor (1783), später Generalpostkontroleur in London, mit Aufhängen versehen und erheblich verbessert, war der Vorläufer der preussischen, nachmals so berühmt gewordenen Schnellposten; aber auch zur Ausbildung

den, Kurjer Bau Gesch. Gene mit T deutsch ziert lung wesen hinum eine Hülf Post

1. Alter Dresdener Potentknecht v. 1573.

mur, ohne weiteres einem Hogarth als Vorwurf zu einem Genrebilde dienen konnte.

Die englische Mail-Coach, von Mr. Palmer, dem Bristol Theaterdirektor (1783), später Generalpostkontroleur in London, mit Aufhängen versehen und erheblich verbessert, war der Vorläufer der preussischen, nachmals so berühmt gewordenen Schnellposten; aber auch zur Ausbildung

Verifizieren: Bildstapel verifizieren [DM_Uni_Würzburg]

Bildstapel Dokument Feld Tabelle Ansicht Extras Hilfe

Bildleiste Bildbereich

Klassifikation

Klasse (Alle): Daheim/Default

Detailbereich

Vorname

Königin von Polen

Arbeitsbereich

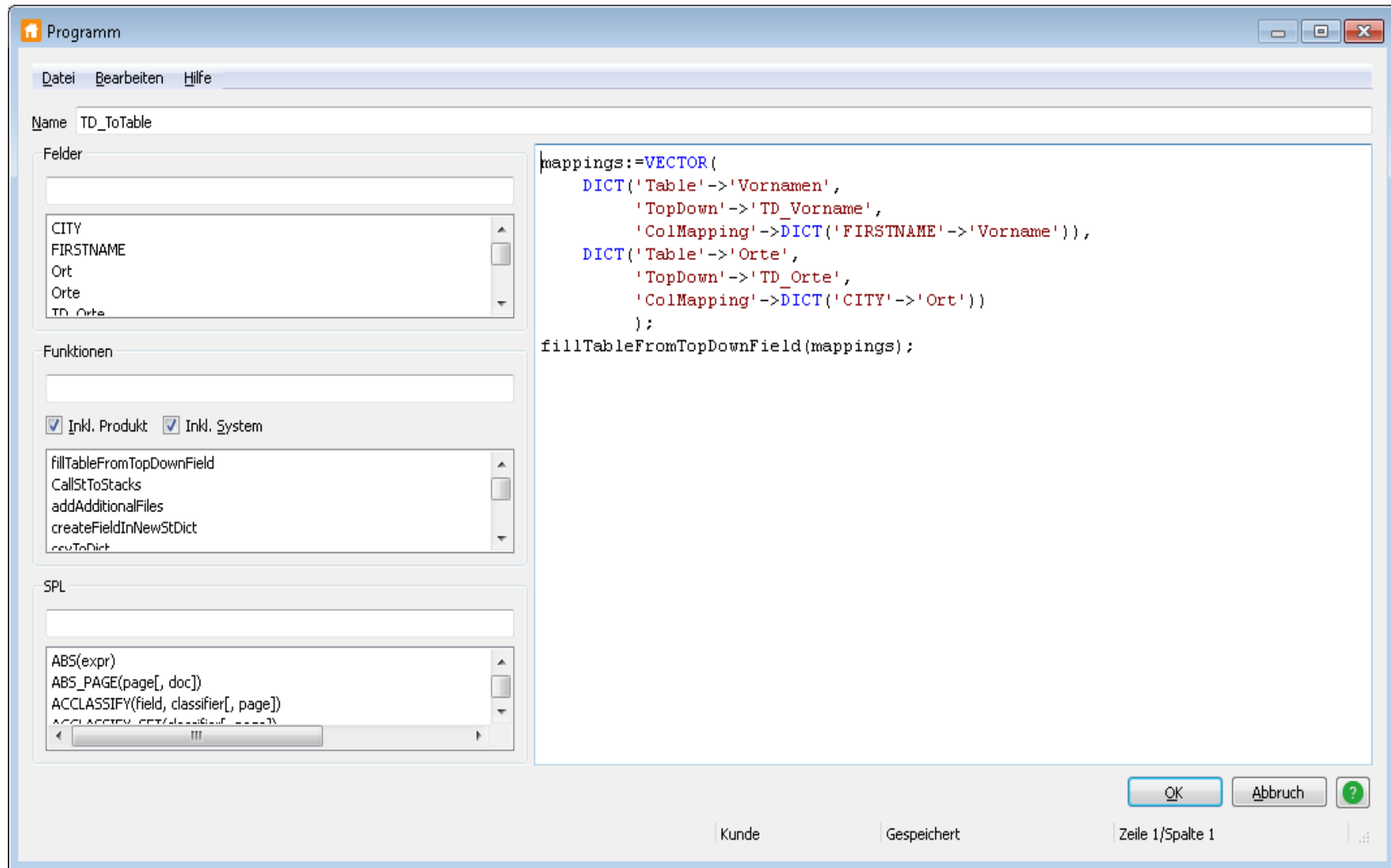
Orte

	Ort
1	Berlin
2	Düsseldorf
3	Merseburg
4	Hagen
5	Halle
6	

Vornamen

	Vorname
1	Elisabeth
2	Friedrich
3	Königin von Polen
4	

Vorgang NONE Dokument 0000 Seite 1/1 Bild 1/1 DE-20_54_A_101-2077-10_27-52_0314_0826_JBH.1874_10_52.pdf:0 VER





Oppertshausen, den 25.8.2016

Kündigung meines Kontos

Sehr geehrte Damen und Herren,

ich würde gern mein Konto bei Ihnen zum Jahresende (30.12.2016) kündigen und auflösen.

Meine Girokontonummer bei Ihnen lautet: 0218 6233 00

Bitte überweisen Sie am 01.01.2017 das verbleibende Guthaben auf folgendes Konto:

Kontoinhaber: Tom Goldschmidt
Kontonummer: 4599 297 00
BLZ: 544 688 00
Kreditinstitut: 24/7 Deutschland Bank

Eventuell ausstehende Verbindlichkeiten können über dieses Konto eingezogen werden.
Eine entsprechende einmalige SEPA-Mandatsverfügung habe ich Ihnen mit beigefügt.

Bitte bestätigen Sie mir den Eingang der Kündigung schriftlich an folgende Adresse (Achtung: abweichend der bei Ihnen hinterlegten Adresse):

Tom Goldschmidt
c/o Nina Müller
Rathausstraße 35,
90215 Nürnberg

Viele Grüße,

Edmund

FREIPOST/KÜNDIGUNG KONTO

SPRACHE

DE | EN | FR | IT

Vertragsende

30.12.2016

KUNDE

p

IBAN

DE13123456780218623300

Vorname | Nachname

Tom Goldschmidt

Strasse

Budapester Straße 68

PLZ | Ort

55469 Oppertshausen

BESCHWERDEERKENNUNG

Ernste Beschwerde

Beschwerde

Beschwerdehinweis

NEUE ADRESSE

Strasse

Rathausstraße 35

PLZ | Ort

90215 Nürnberg

Telefon

E-Mail-Adresse

GUTHABENERSTATTUNG

Kontoinhaber

Tom Goldschmidt

IBAN

DE21 5446 8800 0459929700

Kreditinstitut

24/7 Deutschland Bank

Unterschrift vorhanden

MELDUNGEN KOMMENTARE

[illegible]

Hardware Shop

...the easy way

Oklahoma Drive 1002
Tulsa, OK 74194
Phone (509) 555-0190 Fax (509) 555-0191

DATE: October 15, 2011
INVOICE # 1457633

Bill to: Insiders Technologies GmbH
Brüsseler Straße 1
67657 Kaiserslautern
Germany

Ship to:

Comments or Special Instructions:

ORDER DATE	P.O. NUMBER	SHIP DATE	DELIVERY NO.	CUSTOMER NO.	SALES PERSON
October 07, 2011	14358962	October 10, 2011	D578934	5786269	Ms Miller

QUANTITY	DESCRIPTION	UNIT PRICE	AMOUNT
2	8968368 Mainboard Intel	\$ 99.00	\$ 198.00
5	2817652 Notebook HP	449.00	2,245.00
3	1435987 Buffalo 3.0 Drive Station	79.90	239.70
4	2456389 Dual Core Opteron	189.90	795.60
1	1020168 External Harddrive 3.0	49.90	49.90

SUBTOTAL \$ 3,492.20

TAX RATE 0.00 %

SALES TAX -

SHIPPING & HANDLING \$ 96.80

TOTAL NET \$ 3,589.00

TOTAL \$ 3,589.00

Make all checks payable to **Hardware Shop**
If you have any questions concerning this invoice, contact Ms Miller m.miller@hardware-shop.com

THANK YOU FOR YOUR BUSINESS!

RECHNUNG/STANDARD

Hardware Shop
Oklahoma Drive 1002
US 74194 Tulsa, OK
Steuernr. USt-IdNr.
Swift-BIC
IBAN

Rechnungsdatum | -Nr. 15.10.2011 1457633

Leistungsdatum

Netto 1 | MwSt. % 1 | MwSt. 1 3.589,00 0,00 0,00

Währung | Betrag USD 3.589,00

Mit Bestellbezug ☒

BESTELLUNGEN

BestellNr.	LS-Nr.	LS-Datum	Betrag
14358962			3.528,20

RECHNUNGSPPOSITIONEN

Pos.	BestellNr.	Menge	Einzelpreis	Einheit	Gesamtpreis
10	14358962	2,00	99,000	1,00	198,00
20	14358962	5,00	449,000	1,00	2.245,00
30	14358962	3,00	79,900	1,00	239,70
40	14358962	4,00	189,900	1,00	795,60

ZAHLUNGSBEDINGUNGEN

Tage	Fälligkeit	Skt.%	Skt. Betr.	Zahlbetr.

Rechnungskorrektur ☐

Mahnung | Stufe ☐

UStG-konform ☒

MELDUNGEN KOMMENTARE

Fehlerhafte Rechnungsposition (759,60 ungleich 795,60)

Positionssumme passt nicht zum Nettobetrag (3478,30 ungleich 3589,00)

HARDWARESHOP

Hardware Shop

OKlahoma Drive 1002
Tulsa, OK 74194
Phone (509) 555-0190 Fax (509) 555-0191

DATE: October 15, 2011
INVOICE # 1457633

Bill to: Insiders Technologies GmbH
Brüsseler Straße 1
67657 Kaiserslautern
Germany

Ship to:

Comments or Special Instructions:

ORDER DATE	P.O. NUMBER	SHIP DATE	DELIVERY NO.	CUSTOMER NO.	SALES PERSON
October 07, 2011	14358962	October 10, 2011	D578934	5786269	Ms Miller

QUANTITY	DESCRIPTION	UNIT PRICE	AMOUNT
2	8968368 Mainboard Intel	\$ 99.00	\$ 198.00
5	2817652 Notebook HP	449.00	2,245.00
3	1435987 Buffalo 3.0 Drive Station	79.90	239.70
4	2456389 Dual Core Opteron	189.90	795.60
1	1020168 External Harddrive 3.0	49.90	49.90

SUBTOTAL \$ 3,492.20

TAX RATE 0.00 %

SALES TAX -

SHIPPING & HANDLING \$ 96.80

TOTAL NET \$ 3,589.00

TOTAL \$ 3,589.00

Make all checks payable to **Hardware Shop**
If you have any questions concerning this invoice, contact Ms Miller m.miller@hardware-shop.com

THANK YOU FOR YOUR BUSINESS!

Hardware Shop

OKlahoma Drive 1002
Tulsa, OK 74194
Phone (509) 555-0190 Fax (509) 555-0191

DATE: October 15, 2011
INVOICE # 1457633

Bill to: Insiders Technologies GmbH
Brüsseler Straße 1
67657 Kaiserslautern
Germany

Ship to:

Comments or Special Instructions:

ORDER DATE	P.O. NUMBER	SHIP DATE	DELIVERY NO.	CUSTOMER NO.	SALES PERSON
October 07, 2011	14358962	October 10, 2011	D578934	5786269	Ms Miller

QUANTITY	DESCRIPTION	UNIT PRICE	AMOUNT
2	8968368 Mainboard Intel	\$ 99.00	\$ 198.00
5	2817652 Notebook HP	449.00	2,245.00
3	1435987 Buffalo 3.0 Drive Station	79.90	239.70
4	2456389 Dual Core Opteron	189.90	795.60
1	1020168 External Harddrive 3.0	49.90	49.90

SUBTOTAL \$ 3,492.20

TAX RATE 0.00 %

SALES TAX -

SHIPPING & HANDLING \$ 96.80

TOTAL NET \$ 3,589.00

TOTAL \$ 3,589.00

Make all checks payable to **Hardware Shop**
If you have any questions concerning this invoice, contact Ms Miller m.miller@hardware-shop.com

THANK YOU FOR YOUR BUSINESS!

Office Supplies International

1435 Main Road
Houston, TX 77058
Phone (281) 455-8888
Fax (281) 455-8888
Email info@office-supplies.com

DATE: October 15, 2011
INVOICE # 1457633
ORDER DATE: 09-29-2011

Bill to: Insiders Technologies GmbH
Brüsseler Straße 1
67657 Kaiserslautern
Germany

Ship to:

Comments or Special Instructions:

ITEM NUMBER	DESCRIPTION	QUANTITY	UNIT PRICE	AMOUNT
1435-001	Office Paper 8 1/2 x 11	10	\$ 1.99	\$ 19.90
1435-002	Ballpoint Pen Blue	100	\$ 0.10	\$ 10.00
1435-003	Stapler Desktop	1	\$ 19.99	\$ 19.99
1435-004	Stapler Desktop	1	\$ 19.99	\$ 19.99

SUBTOTAL \$ 69.88

TAX RATE 0.00 %

SALES TAX -

SHIPPING & HANDLING \$ 96.80

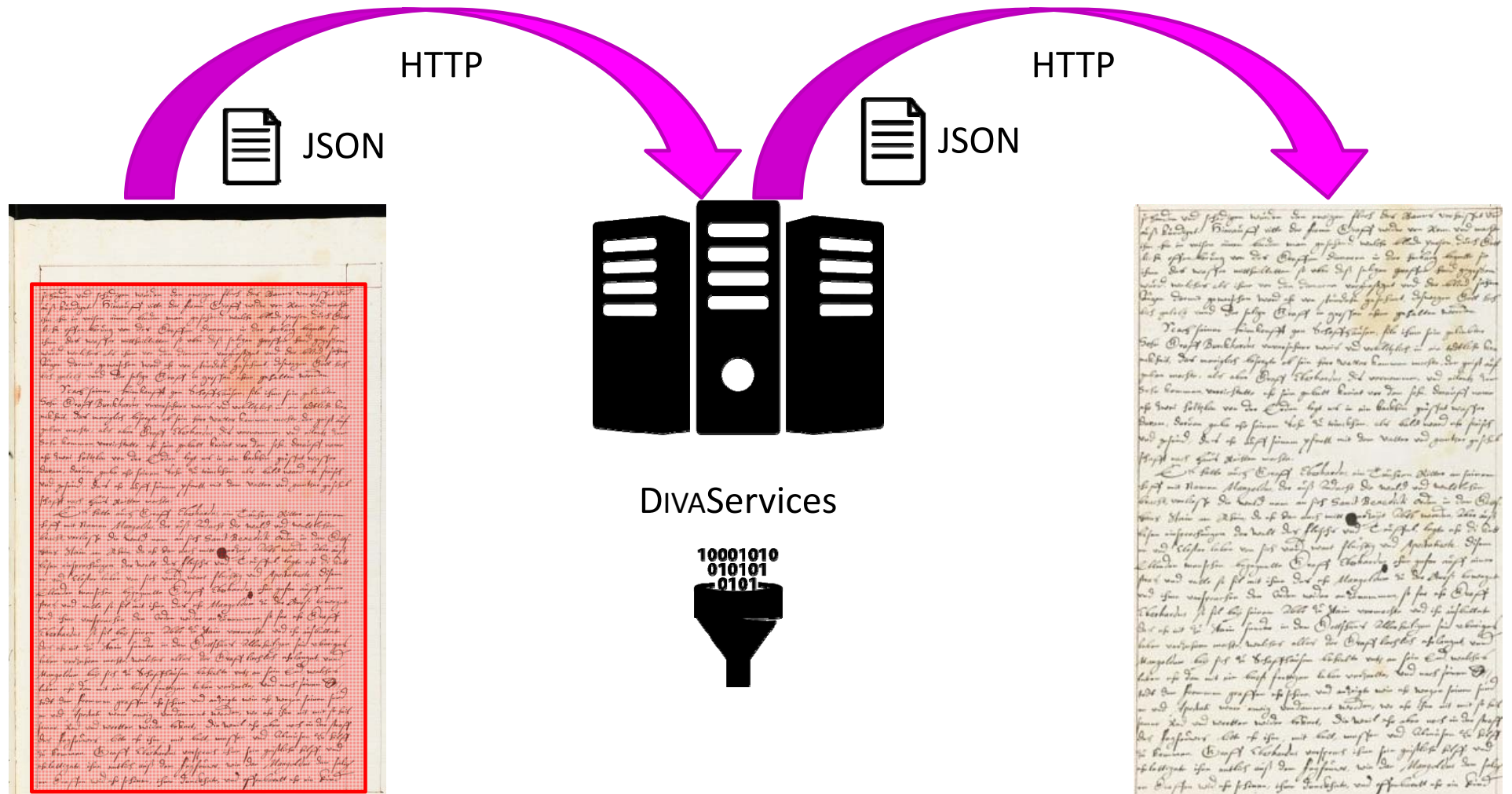
TOTAL NET \$ 166.68

TOTAL \$ 166.68

Make all checks payable to **Office Supplies International**
If you have any questions concerning this invoice, contact Ms Miller m.miller@office-supplies.com

THANK YOU FOR YOUR BUSINESS!

Oder alternativ: Open DIVAServices



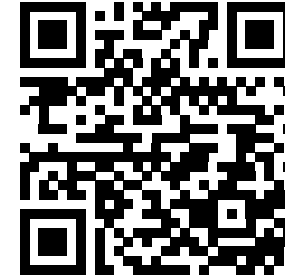
Client is anywhere on the world

MindGarage DIVAServices & Spotlight

- Services sind OpenSource
- Bilder alle unter CreativeCommons
- Docker-Technologie für Methoden

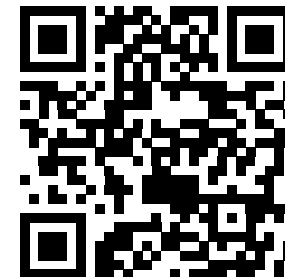
Spotlight

- DAS 2016 Best Paper Award
- Bereits in diversen Tools in Fribourg und anderen Unis verwendet
- Verwendet in Wettbewerben
- Workshops (ICDAR-OST) und Summer Schools



Project Website

<http://bit.ly/divaservices>



DivaServices-Spotlight

<http://divaservices.unifr.ch/spotlight>

- Deep Learning basiert auf Simulation von menschlichen Neuronen
- Methoden sind state-of-the-art in Binärisierung, Layout Analyse, OCR, Post-Processing

Zwei Möglichkeiten der Integration

- Kommerziell
 - ^ Vorteile: nachhaltig, anpassbar, generisch, umfassend, praktische Erfahrung
 - ^ Nachteile: Kosten, Abhängigkeit(?), Kontrolle
 - ^ Möglichkeit: OpenSource, aber Anbieten von Services und Erweiterungen
- Forschungstools (z. B. DIVAServices)
 - ^ Vorteile: OpenSource, OpenService, leicht anpassbar
 - ^ Nachteile: nicht nachhaltig, nicht umfassend,

Deep Learning für Automatische Dokumentanalyse

apl. Prof. Marcus Liwicki

DIVA Group, University of Fribourg

MindGarage, University of Kaiserslautern

marcus.liwicki@unifr.ch

www.mindgarage.de